DOCUMENT RESUME

ED 392 844 TM 024 692

AUTHOR	Gershon, Richard; Bergstrom, Betty
TITLE	Does Cheating on CAT Pay: NOT!
PUB DATE	Apr 95
NOTE	23p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*ıdaptive Testing; Algorithms; *Cheating; *Computer Assisted Testing; *Difficulty Level; Estimation (Mathematics); *Item Banks; Simulation; *Test Items
IDENTIFIERS	Ability Estimates; Rasch Model

ABSTRACT

When examinees are allowed to review responses on an adaptive test, can they "cheat" the adaptive algorithm in order to take an easier test and improve their performance? Theoretically, deliberately answering items incorrectly will lower the examinee ability estimate and easy test items will be administered. If review is then allowed, examinees can change answers from wrong to right, thereby raising their initial ability estimate. Following this strategy, examinees can take an easy test, rather than a test targeted to their ability. The consequences of following such a strategy, for the examinee and for the testing agency, are explored. Results of a simulation using data based on the Rasch model indicate that cheating is a risky business. If an examinee makes a mistake and fails to change even one answer from wrong to right, the consequences may be dire. When an item bank has very easy items and test length is short, highly able examinee ability is severely underestimated. In addition, "cheating" can be detected and prevented during test administration by altering test targeting. (Contains 4 figures, 4 tables, and 22 references.) (Author/SLD)

Does Cheating on CAT Pay: NOT!

Richard Gershon

Computer Adaptive Technologies, Inc.
Northwestern University

Betty Bergstrom

Computer Adaptive Technologies, Inc.

Paper presented at the annual meeting
American Educational Research Association Meeting

San Francisco, 1995

Does Cheating on CAT Pay: NOT!

Richard Gershon
Computer Adaptive Technologies, Inc.
Northwestern University

Betty Bergstrom
Computer Adaptive Technologies, Inc.

Abstract

When examinees are allowed to review responses on an adaptive test, can they "cheat" the adaptive algorithm in order to take an easier test and improve their performance? Theoretically, deliberately answering items incorrectly will lower the examinee ability estimate and easy test items will be administered. If review is then allowed, examinees can change answers from wrong to right thereby raising their initial ability estimate. Following this strategy, examinees can take an easy test, rather than a test targeted to their ability. The consequences of following such a strategy, for the examinee and for the testing agency, are explored. Results indicate that cheating is a risky business. If an examinee makes a mistake, and fails to change even one answer from wrong to right, the consequences may be dire.. When an item bank has very easy items and test length is short, high able examinee ability is severely underestimated. In addition, "cheating" can be detected and prevented during test administration by altering test targeting.

Does Cheating on CAT Pay: NOT!!

When examinees are allowed to review responses on an adaptive test, can they "cheat" the adaptive algorithm in order to take an easier test and improve their performance? Theoretically, deliberately answering items incorrectly will lower the examinee ability estimate and easy test items will be administered. If review is then allowed, examinees can change answers from wrong to right thereby raising their initial ability estimate. Following this strategy, examinees can take an easy test, rather than a test targeted to their ability. The consequences of following such a strategy, for the examinee and for the testing agency, is an issue that has been discussed at several recent AERA/NCME sessions (Wise, Johnson, Plake and Gullett, 1990; Wang and Wingersky, 1992; Vispoel, Wang, del la Torre, Bleiler and Dings, 1992; Lunz, Stahl and Bergstrom, 1993), but has not been systematically explored.

Opponents of allowing examinee review and answer changing on adaptive tests argue that altering responses compromises the efficiency of the adaptive algorithm (Wainer, 1993). Proponents of review contend that examinees like the security of being able to review items and that the number of items they actually change is few and does not substantially alter the precision of measurement (Lunz, Bergstrom and Wright, 1992.)

Many studies have addressed the issue of review and changing answers on paper and pencil examinations. The literature consistently supports the assertion that people who change answers tend to gain. Schwarz, McMorris and DeMers (1991; McMorris, DeMers and Schwarz, 1987) have extensively studied answer changing behavior and report that students gain from changing answers, with students in the upper two thirds of classes improving the most. They also note that students reported changing answers for "thoughtful reasons such as rereading, rethinking, or remembering more information." They caution that the role of memory should be considered in understanding the impact of answer changing.

Benjamin, Cavell and Shallenberger (1984) reviewed 33 studies on changing answers on paper and pencil tests. They recount that "after more than a half century of research on this topic" the evidence uniformly indicates that a) only a small percentage of answers are actually changed, b) the majority of answers are changed from wrong to right, c) most test takers are answer changers, and d) most answer changers are point gainers.

Computerized Adaptive Testing

On a computerized adaptive test (CAT), each examinee takes an individualized test comprised of items chosen from a content validated item bank and tailored to his ability. In general, when an examinee answers an item correctly, the on-line estimation of his ability increases and the next item administered is more difficult. Similarly, when an examinee answers an item incorrectly, the on-line estimation of his ability is lowered and the next item administered is easier.

Computerized adaptive testing has been embraced by certification and licensure organizations because the adaptive process maximizes the precision of measurement, thus allowing test length to be shortened, and increases test security by presenting individualized tests (Bergstrom and Lunz, 1992; National Council State Boards of Nursing, 1993). These CATs are pass/fail high stakes tests and the stopping rule implemented is based on a fixed number of items or on a specified level of confidence in the pass/fail decision. Computerizeo adaptive testing is also being used to reduce testing time without decreasing measurement precision for high stakes achievement tests sucli as the Graduate Record

Examination (Reese, 1992). High stakes achievement tests typically report a scaled score and the stopping rule is based on a fixed number of items or a fixed level of precision. CATs are also being used as diagnostic tests for course placement in college and university settings (Legg and Buhr, 1987; Doucette, D., 1988; ACT, 1994) and for diagnostic placement in high schools and elementary schools (Kingsbury, 1990; Baghi, Gabrys, Ferrara, 1991). Again, scores are typically reported as scaled scores and the stopping rule implemented is based on a fixed number of items or a specified level of precision.

Allowing Review on a CAT

In general, review has not been allowed on adaptive tests. To our knowledge, the only currently administered large scale CATs allowing review are the certification examinations delivered by the Board of Registry (BOR) of the American Society of Clinical Pathologists. Review has been extensively studied by the BOR (Lunz, Bergstrom and Wright, 1992; Lunz and Bergstrom, 1994 and Lunz and Bergstrom, In Press) and is allowed on all 17 of their certification tests (Gershon, 1994). Studies by the BOR on review have strongly replicated findings from paper and pencil studies. Examinees were found to change few answers, more answers were changed from wrong to right than from wrong to wrong or from right to wrong, and examinees slightly improved their performance by changing answers. These studies also showed that the effect of changing answers on the precision of measurement was minimal. Under actual certification testing conditions, 70% of the examinees elected to change some responses and the mean increase in SEM was only .002 logits (Lunz et.al., In Press). This information "loss" could be recovered by increasing overall test length by just one item.

However, concerns have been raised that examinees might use the review process to "cheat" on an adaptive exam. On a paper and pencil test, all examinees receive a fixed set of questions. On a CAT, however, the items administered depend upon the performance of the examinee. Examinees who perform poorly will be administered easy items. The concern is that examinees will deliberately use this process to obtain a test comprised of easy items and then during review answer the items correctly. In this paper we will examine:

1) The consequences of attempted cheating on the examinee ability estimate and standard error of measure (SEM).

2) The effect of item bank width on the consequences of attempted cheating.

3) The effect of test length on the consequences of attempted cheating.

Method

Data shown in this paper are based on the Rasch model and obtained by using the PROX estimation method (Wright and Stone, 1979). In order to purposefully answer items incorrectly, examinees must be smart enough to know the right answer and then deliberately choose a wrong answer. Therefore the number of questions that an examinee can purposefully miss can be predicted by the Rasch model. The PROX formula estimates person ability with the formula:

$$B_n=\overline{D_i}+LOG(R/(L-R))$$

where: B_n is the estimate of person ability; $\overline{D_i}$ is the mean difficulty of items presented; R is the number of items answered correctly and L is the test length. For this study, in order to mimic the effect of cheating, we set the mean difficulty of items presented to -2.00 logits (easy items) and the mean difficulty of items presented to -4.00 logits (very easy items).

Across a range from -3.5 logits to +3.5 logits at .50 logit intervals, we simulated examinees who attempted to incorrectly answer all of the items on a test and then answer them correctly during review. The probability of correct response, estimated number correct, estimated ability, and standard error of measure were calculated for test lengths of 30 and 90 items, and for banks with easy items and very easy items. Estimates were calculated under the assumption that the examinee was able to purposefully incorrectly answer the model expected number of items (based on ability and item difficulty) and then answer them correctly during review. Of course, the capability of examinees to implement this strategy is modified by their ability relative to item difficulty--a more able examinee is more successful at this strategy than a low able examinee.

Results

Figures 1 to 4 show true ability versus ability estimated after review for tests of 90 and 30 items when the mean item difficulty is set at -2.00 logits and -4.00 logits. Tables 1 and 4 present conversion tables of raw scores to logit measures with their associated standard errors of measure (Wright and Linacre, 1993). These figures and graphs illustrate the effect of deliberately taking an easy test on ability estimation. While examinees are measured comparably regardless of the length of test or mean difficulty of items presented, ceiling effects result in severe underestimation of the ability of high able examinees. Administration of easy or very easy items eliminates the capacity of the test to differentiate between levels of ability at the upper ends of the scale (notice the staircase effect of ability estimation on Figures 1 to 4, especially for the 30 item tests).

For tests of 90 items, the maximum ability estimate obtainable when an easy test (mean item difficulty = -2.00) is administered is 3.19 logits (see Figure 1 and Table 1). However when test length is shortened to 30 items, the maximum ability estimate obtainable is only 2.08 logits (Figure 2 and Table 2.)

For a very easy test (mean item difficulty = -4.00 logits) of 90 items, the maximum ability estimate is 1.19 logits (Figure 3 and Table 3.) The ceiling is even more severe when test length is short. For a test of 30 items, (Figure 4 and Table 4), the maximum ability estimate is only .08 logits.

Tables 1 to 4 show that cheating is a highly risky venture. During review, in order to successfully cheat, the examinee must correctly answer *all* of the items that he purposefully missed. If an examinee makes a mistake, and fails to change even 1 answer from wrong to right, the consequences may be dire. When an item bank has very easy items and test length is short, examinee ability is most severely underestimated if the examinee makes even one or two mistakes. For example, as shown in Table 4, examinees with high ability would be expected to answer all 30 items correctly. However, the estimate of ability drops more than half a logit, from .08 to -.63, if they miss just one of the 30 items and to -1.63 if they miss two of the 30 items.

The effect of the cheating strategy on the standard error of measure (SEM) can also be seen in Tables 1 to 4. The more able the candidate, the greater number of items he will be able to purposefully answer incorrectly and then correct in review. The more the

TEST LENGTH = 90 MEAN ITEM DIFFICULTY = -2.00

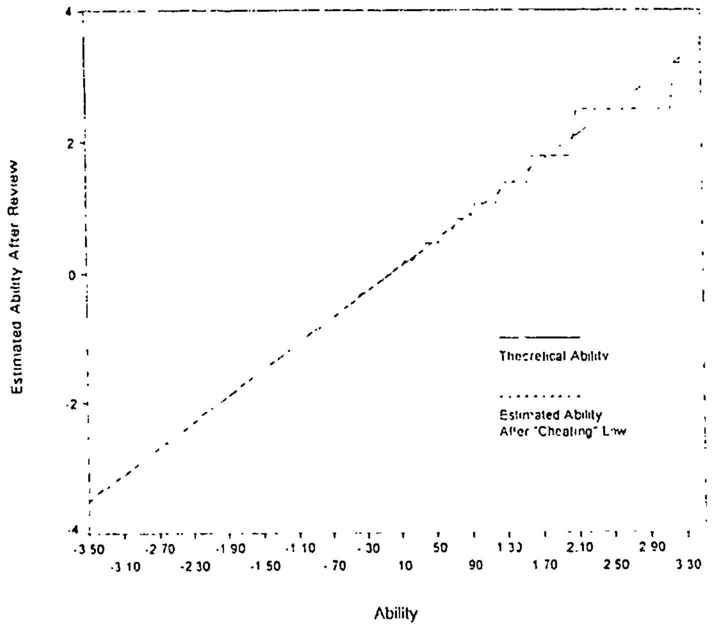

Figure 1. Test Length = 90; Mean Item Diff = -2.00 Comparison of true ability with ability estimated after review

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	-7.19E	1.42	31	-2.64	.22	62	-1.21	.23
1	-6.49	1.01	32	-2.59	.22	63	-1.15	.23
2	-5.78	.72	33	-2.55	.22	64	-1.10	.23
3	-5.37	.59	34	-2.50	.22	65	-1.04	.24
4	-5.07	.51	35	-2.45	.22	66	-.99	.24
5	-4.83	.46	36	-2.41	.22	67	-.93	.24
6	-4.64	.42	37	-2.36	.21	68	-.87	.25
7	-4.47	.39	38	-2.31	.21	69	-.81	.25
8	-4.33	.37	39	-2.27	.21	70	-.75	.25
9	-4.20	.35	40	-2.22	.21	71	-.68	.26
10	-4.08	.34	41	-2.18	.21	72	-.61	.26
11	-3.97	.32	42	-2.13	.21	73	-.54	.27
12	-3.87	.31	43	-2.09	.21	74	-.47	.28
13	-3.78	.30	44	-2.04	.21	75	-.39	.28
14	-3.69	.29	45	-2.00	.21	76	-.31	.29
15	-3.61	.28	46	-1.96	.21	77	-.22	.30
16	-3.53	.28	47	-1.91	.21	78	-.13	.31
17	-3.46	.27	48	-1.87	.21	79	-.03	.32
18	-3.39	.26	49	-1.82	.21	80	.08	.34
19	-3.32	.26	50	-1.78	.21	81	.20	.35
20	-3.25	.25	51	-1.73	.21	82	.33	.37
21	-3.19	.25	52	-1.69	.21	83	.47	.39
22	-3.13	.25	53	-1.64	.21	84	.64	.42
23	-3.07	.24	54	-1.59	.22	85	.83	.46
24	-3.01	.24	55	-1.55	.22	86	1.07	.51
25	-2.96	.24	56	-1.50	.22	87	1.37	.59
26	-2.90	.23	57	-1.45	.22	88	1.78	.72
27	-2.85	.23	58	-1.41	.22	89	2.49	1.01
28	-2.79	.23	59	-1.36	.22	90	3.19E	1.42

Table 1. Test Length = 90; Mean Item Diff = -2.00 Table of measures on complete test (Wright and Linacre, 1993)

TEST LENGTH = 30 MEAN ITEM DIFFICULTY = -2.00

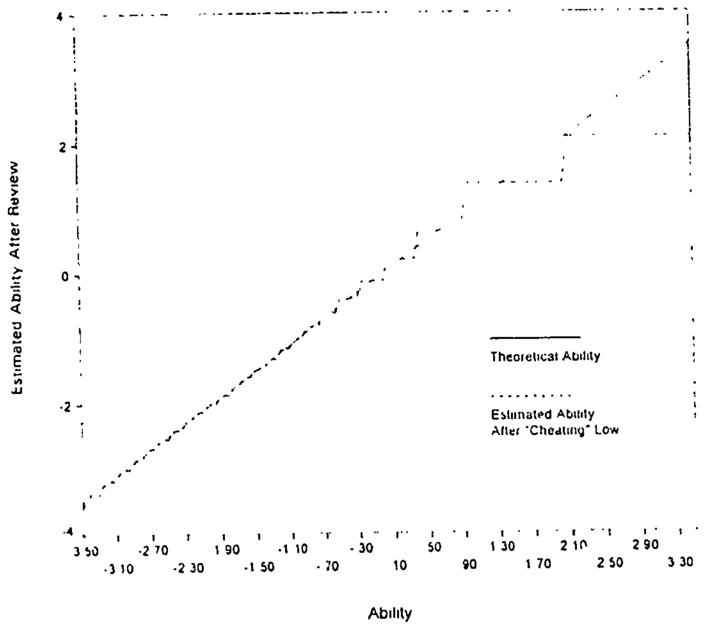

Figure 2. Test Length = 30; Mean Item Diff = -2.00; Comparison of true ability with ability estimated after review

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	-6.08E	1.43	11	-2.55	.38	22	-.99	.41
1	-5.37	1.02	12	-2.41	.37	23	-.81	.43
2	-4.64	.73	13	-2.27	.37	24	-.61	.46
3	-4.20	.61	14	-2.13	.37	25	-.39	.49
4	-3.87	.54	15	-2.00	.37	26	-.13	.54
5	-3.61	.49	16	-1.87	.37	27	.20	.61
6	-3.39	.46	17	-1.73	.37	28	.64	.73
7	-3.19	.43	18	-1.59	.37	29	1.37	1.02
8	-3.01	.41	19	-1.45	.38	30	2.08E	1.43
9	-2.85	.40	20	-1.31	.39			
10	-2.69	.39	21	-1.15	.40			

Table 2. Test Length = 30; Mean Item Diff = -2.00 Table of measures on complete test (Linacre and Wright, 1993)

TEST LENGTH = 90 MEAN ITEM DIFFICULTY = -4

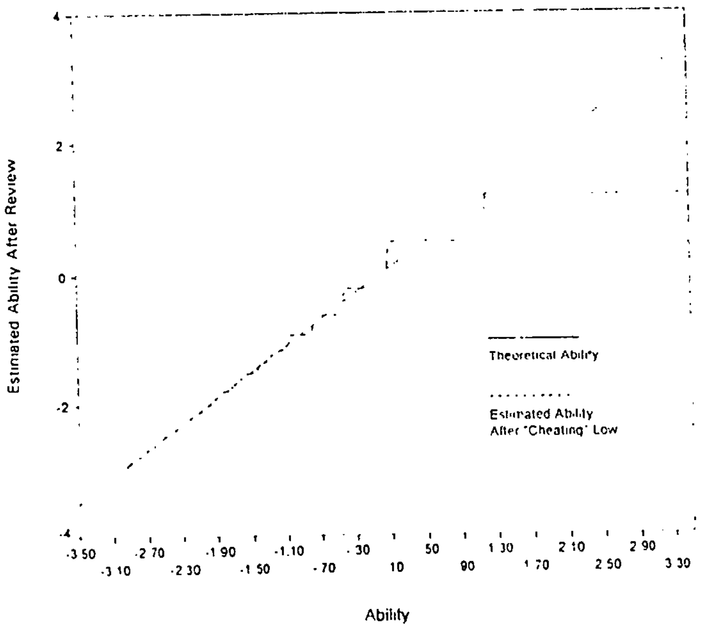

Figure 3. Test length = 30; Mean Item Diff = -2.00 Comparison of true ability with ability estimated after review

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	-9.19E	1.42	31	-4.64	.22	62	-3.21	.23
1	-8.49	1.01	32	-4.59	.22	63	-3.15	.23
2	-7.78	.72	33	-4.55	.22	64	-3.10	.23
3	-7.37	.59	34	-4.50	.22	65	-3.04	.24
4	-7.07	.51	35	-4.45	.22	66	-2.99	.24
5	-6.83	.46	36	-4.41	.22	67	-2.93	.24
6	-6.64	.42	37	-4.36	.21	68	-2.87	.25
7	-6.47	.39	38	-4.31	.21	69	-2.81	.25
8	-6.33	.37	39	-4.27	.21	70	-2.75	.25
9	-6.20	.35	40	-4.22	.21	71	-2.68	.26
10	-6.08	.34	41	-4.18	.21	72	-2.61	.26
11	-5.97	.32	42	-4.13	.21	73	-2.54	.27
12	-5.87	.31	43	-4.09	.21	74	-2.47	.28
13	-5.78	.30	44	-4.04	.21	75	-2.39	.28
14	-5.69	.29	45	-4.00	.21	76	-2.31	.29
15	-5.61	.28	46	-3.96	.21	77	-2.22	.30
16	-5.53	.28	47	-3.91	.21	78	-2.13	.31
17	-5.46	.27	48	-3.87	.21	79	-2.03	.32
18	-5.39	.26	49	-3.82	.21	80	-1.92	.34
19	-5.32	.26	50	-3.78	.21	81	-1.80	.35
20	-5.25	.25	51	-3.73	.21	82	-1.67	.37
21	-5.19	.25	52	-3.69	.21	83	-1.53	.39
22	-5.13	.25	53	-3.64	.21	84	-1.36	.42
23	-5.07	.24	54	-3.59	.22	85	-1.17	.46
24	-5.01	.24	55	-3.55	.22	86	-.93	.51
25	-4.96	.24	56	-3.50	.22	87	-.63	.59
26	-4.90	.23	57	-3.45	.22	88	-.22	.72
27	-4.85	.23	58	-3.41	.22	89	.49	1.01
28	-4.79	.23	59	-3.36	.22	90	1.19E	1.42
29	-4.74	.23	60	-3.31	.22			
30	-4.69	.22	61	-3.26	.23			

Table 3. Test length = 90; Mean Item Diff = -4.00 Table of measures on complete test (Wright and Linacre, 1993)

TEST LENGTH = 30 MEAN ITEM DIFFICULTY = -4

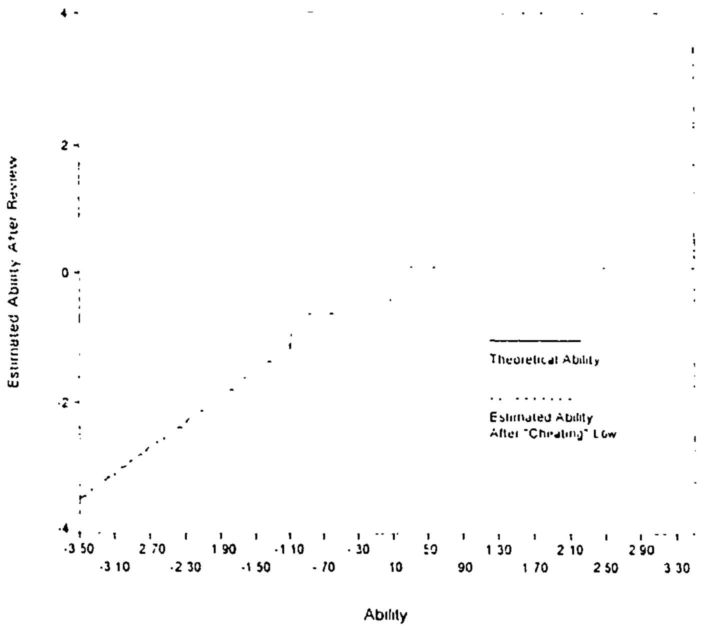

Figure 4. Test Length = 30; Mean Item Diff = -4.00 Comparison of true ability with ability estimated after review

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
0	-8.08E	1.43	11	-4.55	.38	22	-2.99	.41
1	-7.37	1.02	12	-4.41	.37	23	-2.81	.43
2	-6.64	.73	13	-4.27	.37	24	-2.61	.46
3	-6.20	.61	14	-4.13	.37	25	-2.39	.49
4	-5.87	.54	15	-4.00	.37	26	-2.13	.54
5	-5.61	.49	16	-3.87	.37	27	-1.80	.61
6	-5.39	.46	17	-3.73	.37	28	-1.36	.73
7	-5.19	.43	18	-3.59	.37	29	-.63	1.02
8	-5.01	.41	19	-3.45	.38	30	.08E	1.43
9	-4.85	.40	20	-3.31	.39			
10	-4.69	.39	21	-3.15	.40			

Table 4. Test length = 30; Mean Item Diff = -4.00 Table of measures on complete test (Wright and Linacre, 1993)

percentage of items correct departs from 50%, the greater the SEM. In Tables 1 and 3, examinees who correctly answer half (45/90) of the items have a SEM of .21 as opposed to examinees who correctly answer 89 or 90 items and have a SEM of 1.01 and 1.42, respectively. In Tables 2 and 4, examinees who correctly answer half (15/30) of the items have a SEM of .37 as opposed to examinees who correctly answer 29 or 30 items and have a SEM of 1.02 and 1.43, respectively. High able examinees who cheat will be administered very easy items, answer most of them correctly in review, and be measured with very poor precision.

Discussion

Allowing review on a CAT

The effects of review have been poorly understood. Psychometricians and test developers who believe that review cannot be allowed because the adaptive process is altered, fail to understand the basics of how IRT works in an adaptive test drawn from a calibrated item bank. Adaptive testing is sometimes explained as if the process were a simple branching technique: if an examinee answers an item incorrectly, an easier item is administered, if an examinee answers an item correctly, a harder item is administered (Wainer, 1993). This premise implies that if an examinee changes an answer in review, the sequence and difficulty of items presented originally is rendered inappropriate. In fact, adaptive testing is a more sophisticated procedure that involves the re-estimation of ability after each item is administered based upon both the difficulty of and response to *all* items previously administered. If an entire test is administered and then an examinee changes answers, the sequence of the changed item is inconsequential to the effect of the change on

SEM, because ability and SEM are re-estimated based on *all* items and responses. If answer changing occurs during test administration, the on-line estimate of examinee ability and SEM will reflect the changed response and mitigate any off targeting effects of changing answers.

On a CAT, items are targeted to the current on-line estimate of examinee ability. Items are usually randomly chosen from a group of well targeted items in the calibrated bank. Thus the concern regarding exactly which items from the bank might have been administered had the examinee chosen their revised answer initially is not crucial as long as estimated ability does not dramatically change. Examinees who use review to change answers that they were initially unsure of, change some answers from wrong to right but also some from right to wrong and some from wrong to wrong. This pattern of review behavior has been shown to have very little impact on the standard error of measure (Lunz, et.al., 1992; Lunz et.al, In Press)

Cheating from the Examinee Viewpoint

These tables and graphs clearly show that the strategy of cheating to get easy questions is an unwise procedure. High able examinees will definitely depress the estimate of their ability. If only pass/fail status is reported, the examinee faces the possibility that the ceiling will be too low, and that answering all of the questions correctly after review will not raise his ability estimate sufficiently to pass the test.

The Effect of Bank Width on Cheating

Certification and licensure banks are typically targeted to a well specified examinee population, bank width is constrained and therefore the difficulty range of items is restricted. This makes it less likely that an examinee will be able to purposefully get a very easy test.

On achievement tests and diagnostic tests, a wide range of examinee ability is often being tested, bank width is greater, and the range of easy items may extend well below a particular examinee's estimated ability. The effects of cheating may prove particularly detrimental to examinees who are attempting to demonstrate their ability on these tests.

Cheating from the Test Developer Viewpoint

Cheating behavior affects the precision of measurement. Therefore if review is allowed on a CAT, test developers don't want examinees to follow a cheating strategy. Examinees who realize the potential risk of cheating will be less likely to do so. If review is allowed, examinees should understand that, while they may be able to take an easier test, they cannot artificially inflate the estimate of their ability and that cheating may very well result in an underestimation of their ability and/or failing the test.

Cheating however, can be detected and prevented during test administration. Testing agencies who allow review are probably wise to include cheating detection behavior in their algorithm. This can be accomplished by monitoring the percentage of items examinees answer correctly. Since tests are targeted to examinee ability, the expected percent correct hovers about 50 percent. Examinees can be monitored continuously on-line to detect unusual responses patterns (e.g. if percentage correct drops below 30% at any point after the first five items are administered). If cheating is detected, appropriate steps to thwart the charlatan can be taken. The adaptive algorithm can be modified to administer items based upon a strategy other than targeting the difficulty of the item to the current on-line ability estimate. On a pass/fail test, suspect examinees could be administered the maximum number of items targeted to the decision point. For diagnostic and achievement tests, examinees could be administered items with difficulty values near the center of the item difficulty distribution or targeted to appropriate age or grade level. Alternatively, their test could be terminated with directions to contact the test proctor for additional instructions.

Conclusion

Should review be prohibited because an examinee might cheat? Should we disallow review because some examinees may attempt to take an easy test (at great risk to their potential score?) Gaining scaled score points and/or moving from fail to pass may be a critical issue from the perspective of the test taker. From a technological perspective cheating can be detected and thwarted. Therefore, if the primary reason for giving a CAT is accurate and precise measurement, and time limits are not an issue, allowing review ensures that the estimate of examinee ability has allowed for thoughtful response.

Author's Address

Send requests for reprints or further information to

Richard C. Gershon
Computer Adaptive Technologies, Inc.
2609 W. Lunt Ave
Chicago, IL 60614.
Tel: (313) 274-3286 FAX: (312)-274-3287
E-mail: rgershon@catinc.com

Reference

ACT, (1994). *COMPASS Computerized Adaptive Placement Assessment and Support Services* [Computer Software] Educational Services Division.

Baghi, H., Gabrys, R., & Ferrara, S. (1991, April). *Applications of computer-adaptive testing in Maryland*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Benjamin, L. T., Cavell T. A., & Shallenburger, W. R. I. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology*, 11(3), 133-141.

Bergstrom, B. A., & Lunz, M. E. (1992). Confidence in pass/fail decisions for computer adaptive and paper and pencil examinations. *Evaluation and the Health Professions*, *15*(4), 453-464.

CAT Question & Answer Reference Guide. (1993). Chicago, IL: National Council of State Boards of Nursing, Inc.

Doucette, D. (Ed.) (1988). *Computerized adaptive testing: The state of the art in assessment at three community colleges*. Laguna Hills, CA: League for Innovation in the Community College.

Gershon, R. C. (1994). *CATSoftware System [computer program]*. Chicago, IL: Computer Adaptive Technologies, Inc.

Kingsbury, G. G. (1990). Adapting adaptive testing: Using the MicroCAT testing in a local school distri t. *Educational Measurement: Issues and Practice*, *9*(2), 3-6.

Legg, S. M., & Buhr, D. (1987, November). *Final report: Feasibility study of a computerized test administration of the CLAST* [(Contract: 5401473-12)]. Institute for student assessment and evaluation. University of Florida.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computer adaptive test administration formats. *Journal of Educational Measurement*, *31*(3), 251-263.

Lunz, M. E., & Bergstrom, B. A. (In Press). Computerized Adaptive Testing: Tracking Candidate Response Patterns. *Journal of Educational Computing Research*.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency in computer adaptive tests. *Applied Psychological Measurement*, *16*(1), 33-40.

Lunz, M. E., Stahl, J. A., & Bergstrom, B. A. (1993, April). *Targeting, test length, test precision and decision accuracy for computerized adaptive tests.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

McMorris, R. F., Demers, L. P., & Schwartz, S. P. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. *Journal of Educational Measurement*, 24(2), 131-143.

Reese, C. (1992). *Development of a computer-based test for the GRE general test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement*, 28(2), 163-171.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992, April). *How review options and administration modes influence scores on computerized vocabulary tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*(1), 15-20.

Wang, M., Wingersky M. (1992, April). *Incorporating post-administration item response revision into a CAT.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wise, S. L., Johnson, P. L., Plake, B. S., & Nebelsick-Gullett, L. (1990). *Providing examinees the opportunity to review items, skip items and change item choices on computerized achievement tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Wright, B. D., and Linacre, J. M. (1993). *BIGSTEPS [computer program].* Chicago, IL: MESA Press.

Wright, B.D. and Stone, M. (1979). *Best test design.* Chicago: MESA Press.